MCL361 Assignment 2

Sanchit | 2021ME21063

Q1: Linear Regression Analysis

Data Generation: Data points were generated from the model

$$y = a + bx + \text{error}$$

where $a = 0.2$ and $b = 0.3$. The x values were integers ranging from 1 to 50, and the errors were drawn independently from a normal distribution with mean 0 and standard deviation 0.5.

Data Splitting and Model Training: The dataset was split into training and testing sets, with 80% of the data used for training and 20% for testing. A linear regression model was trained using the scikit-learn library. The slope and intercept of the learned model were printed. A scatterplot of the data, along with the fitted regression line, was created. Additionally, a bar plot was produced showing the true y-values and the predicted y-values for each x from the testing set.

Performance Metrics: Using the learned model, predictions for the testing set were made. The following metrics were computed:

- **Mean Absolute Error (MAE):**

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

where y_i represents the true values and $\hat{y}_i$ represents the predicted values.

- **Mean Squared Error (MSE):**

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **Root Mean Squared Error (RMSE):**

$$\text{RMSE} = \sqrt{\text{MSE}}$$

- **Coefficient of Determination (R^2):**

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- **Explained Variance Score (EVS):**

$$\text{EVS} = \frac{\text{Var}(\hat{y})}{\text{Var}(y)}$$

where VAR represents the variance of the values.

Results: The following values were obtained with a random seed of 0:

- **Slope (b):** 0.2885
- **Intercept (a):** 0.5236
- **Mean Absolute Error (MAE):** 0.3633

- **Mean Squared Error (MSE):** 0.1907
- **Root Mean Squared Error (RMSE):** 0.4366
- **Coefficient of Determination ($\mathbf{R^2}$):** 0.9856
- **Explained Variance Score (EVS):** 0.9885

Figures:

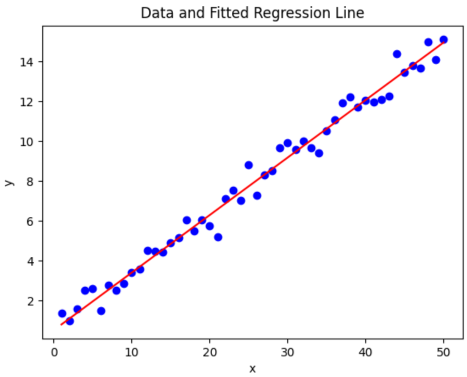

Figure 1: Scatterplot of data with the fitted regression line.

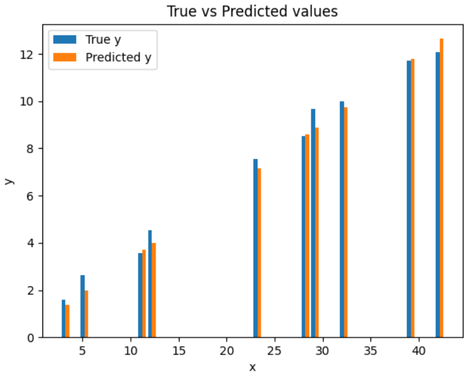

Figure 2: Bar plot of true y-values versus predicted y-values for each x in the testing set.

Q2: Residual Error Analysis

Data Generation and Model Training: Data points were generated from the model

$$y = a + bx + \text{error}$$

with $a = 0.2$ and $b = 0.3$, where x values ranged from 1 to 5000, and errors were drawn from a normal distribution with mean 0 and standard deviation 0.5. The dataset was split into 80% training and 20% testing sets. A linear regression model was trained and used to make predictions for the testing set.

Residual Analysis: The residuals were computed as

$$\text{Residuals} = y_{\text{test}} - \hat{y}_{\text{test}}$$

A histogram of the residuals was plotted. The mean and standard deviation of the normal distribution that best fits the residuals were calculated and printed. The histogram included a plot of the fitted normal distribution to visualize how well it matched the residual distribution.

Results: The following values were obtained with a random seed of 0:

- **Mean of Residuals:** 0.00
- **Standard Deviation of Residuals:** 0.49

Figures:

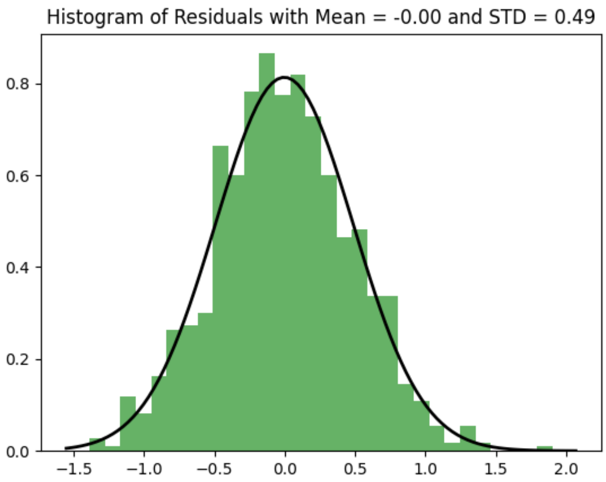

Figure 3: Histogram of residuals with a normal distribution fit.

Q3: OLS Regression Analysis

Data Generation and OLS Fitting: Data points were generated from the model

$$y = a + bx + \text{error}$$

with $a = 0.2$ and $b = 0.3$, where x values ranged from 1 to 50. An Ordinary Least Squares (OLS) regression model was fitted using the statsmodels library.

Slope Checking: The true slope b was checked against the estimated slope from the OLS model:

- The fraction of times the true slope was within one standard error of the estimated slope was calculated.
- The fraction of times the true slope was within two standard errors of the estimated slope was calculated.

The entire process of data generation, model fitting, and slope checking was repeated 1000 times. The reported fractions were compared with the expected values, and an analysis of how close the results were to the expected values was provided.

Expected Values: For a normally distributed error term, it is expected that approximately 68% of the estimated slopes should be within one standard error (SE) of the true slope, and about 95% should be within two SEs of the true slope. Therefore, the expected fractions are approximately 0.68 for one SE and 0.95 for two SEs.

Comparison:

- **Fraction within one SE:** 0.705
- **Fraction within two SEs:** 0.952

The results are close to the theoretical values, with the fraction within one SE being slightly higher than expected and the fraction within two SEs aligning well with the expected 0.95. This indicates that the model's estimates are reasonably accurate.

Q4: Quadratic Model Analysis

Data Generation and Model Fitting: Data points were generated from the model

$$y = a + bx + cx^2 + \text{error}$$

with $a = 3$, $b = 8$, and $c = 20$. The x values were randomly sampled 100 times from a uniform distribution in the range [0, 50], and errors were drawn from a normal distribution with mean 0 and standard deviation 3.

Model Comparison: Both a linear and a quadratic model were fitted to the data using OLS. The Akaike Information Criterion (AIC) values for both models were compared to determine which model fit the data better. The quadratic model was expected to provide a better fit due to its ability to capture the quadratic relationship. The coefficients a, b, and c were evaluated to determine if they were accurately recovered using the quadratic model fitting.

Model Comparison Using AIC:

- **AIC for Linear Model:** 1936.36
- **AIC for Quadratic Model:** 506.83

Justification: The Akaike Information Criterion (AIC) values indicate that the quadratic model has a much lower AIC compared to the linear model. A lower AIC suggests a better fit of the model to the data, accounting for the complexity of the model. Hence, the quadratic model is preferred over the linear model.

Coefficient Recovery: The coefficients of the quadratic model were found to be approximately:

- $a \approx 2.854$
- $b \approx 8.101$
- $c \approx 19.998$

These coefficients are close to the true values used in the data generation ($a = 3$, $b = 8$, $c = 20$), showing that the quadratic model was effective in recovering the true parameters.

Figures:

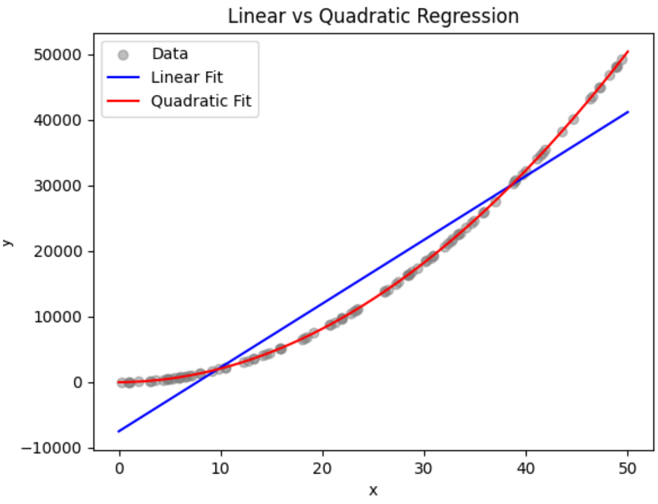

Figure 4: Quadratic model fit to the data.